Niveau(x) d'enseignement: 7 Numéro EDUQ : 0479091
Centre d'intérêt: 570

Hamel, Claude, 1944-.

Test PERPE général : validité / Claude Hamel. -- [Montréal : PERPEI, 1972.
35, [1] f. ; 28 cm.

Document: 0672-03. Bibliogr.: f. [33-36].

Descripteur(s): *Evaluation de l'enseignant *Evaluation par l'étudiant *Satisfaction de l'étudiant *Attitude de l'enseignant *Relations maître-élèves *Questionnaire *Evaluation de tests *Test *Validité *Collégial (Niveau)

Identificateur(s) de nom: *PERPE (Test)

La première partie de ce texte présente diverses conceptions du problème de la validité d'un instrument de mesure, qui détermine jusqu'à quel point celui-ci répond aux objectifs de son utilisation. Elle traite de la validité de contenu, de la validité prédictive et concomitante et de la validité de construit. La deuxième partie prend l'allure d'un document de travail qui porte sur la validation du test PERPE (Perceptions étudiantes de la relation professeur-étudiants) et traite des aspects suivants: description du test PERPE collégial, problème des objectifs visés par l'utilisation de ce test, validité de contenu du test, validité prédictive et concomitante, validité de construit, où on propose une stratégie en cinq étapes.

Type de document: r

EDUQ 8470091

60

Test PERPE Général

Validité

Claude Hamel

Juin 1972

Document 0672-03

Validité

C'est l'une des caractéristiques les plus importantes puisqu'elle détermine l'assurance avec laquelle l'usager utilisera un instrument. Cette caractéristique permet de déterminer jusqu'à quel point l'instrument de mesure répond aux objectifs de son utilisation. Selon la nature de ces objectifs, différents types de validité seront étudiés. Quatre types de validité sont habituellement associés à ces différents objectifs: validité de prédiction, validité concomittante, validité de construit, validité de contenu. Nous allons, dans une première partie, expliciter ces différents types de validité, nous allons étudier les techniques de validation appropriées au test PERPE.

Auparavant, il peut être utile d'isoler quelques éléments de comparaisons entre le concept de validité et celui de fidélité. Cureton (16), pour sa part, considère la fidélité comme un aspect particulier du problème de la validité. La validité comporte deux aspects, la pertinence et la fidélité. La pertinence d'un test a rapport au degré d'accord entre les mesures effectuées dans le test et la fonction que le test vise à mesurer; la fidélité concerne la consistence de cette mesure dans l'échantillon utilisé. Cette conception a quelque intérêt si l'on considère les liens étroits unissant ces deux aspects. En effet, un test pour être valide, du moins quantativement, doit d'abord être fidèle (Lord (37), page 72). Il faut ajouter qu'un test fidèle n'est pas nécessairement valide. La fidélité est donc une condition nécessaire, mais non suffisante, à la validité d'un test. Cette antériorité, par contre, justifie amplement l'importance que les auteurs donnent à l'analyse de la fidélité comme telle.

Le deuxième élément de comparaisons concerne l'aspect "précision" d'une mesure. Ebel (18) tout en adoptant le principe des deux aspects liés au concept de validité, définit l'aspect "fidélité" en termes de précision d'une mesure. Les techniques et la méthodologie qu'il propose vise par contre à évaluer la consistence et la stabilité temporelle des mesures d'un test. A notre avis, la précision d'une mesure concerne exclusivement le

rapport entre une mesure observée et une mesure-étalon; en tant que telle, elle appartient à l'aspect "pertinence" d'une mesure.

C'est l'aspect "pertinence" qui retiendra notre attention dans ce document de travail et c'est en adoptant la terminologie proposée par l'Association Américaine de Psychologie (3) que nous entreprendrons sa discussion.

I - Validité de contenu

Ce n'est que depuis quelques années que le terme 'validité de contenu' est couramment employé par les divers auteurs. Auparavant, on parlait de validité empirique ou statistique (Lennon (34), Lord (37)), de validité logique (Lennon (34)), de validité de face (Anastasi (4)), de validité intrinsèque (Goodenough (27))...etc.. Cette polyvalence de termes reflétait deux conceptions différentes de ce type de validité, conceptions identifiées par Lennon (34) comme empirique versus logique et par Lord (37) comme empirique versus théorique.

La conception empirique de la validité de contenu met l'accent sur la relation entre une série d'item et un critère global. L'appellation "empirique" vient du fait que les item choisis le sont sur une base expérimentale, à savoir leur corrélation avec le critère global. Ce type de validité s'insère donc à l'étape de la construction du test et plus particulièrement à celle de l'analyse d'item. La technique la plus pertinente pour effectuer ce choix est la technique de régression multiple. Les item conservés par cette méthode d'analyse auront de fortes corrélations avec le critère et de faibles corrélations entre eux. Le peu d'importance accordée à l'homogénéité des item fera qu'habituellement le test mesure plusieurs choses. Ce type de validité, grandement statistique, prête donc peu d'attention à la pertinence logique du contenu des item.

Par contre, la conception logique ou théorique de la validité de contenu base sa méthodologie sur la relation entre les item choisis et ce qui est mesuré par le test. Les adeptes de cette approche s'appuient sur

une théorie psychologique pour définir un ou plusieurs concepts théoriques qu'ils veulent mesurer à l'aide d'un test. Cette optique conduit à choisir pour un test des item qui mesurent le même trait ou les mêmes traits. L'accent est donc mis ici sur l'homogénéité des item. La présence de cette homogénéité ne suffit pas, par contre, à affirmer que les item mesurent bien le trait désiré, ce n'est qu'en étudiant la validité de construit que cette question sera solutionnée; cependant l'homogénéité donne un indice sérieux sur le fait que tous les item ont quelque chose en commun, on s'attendra donc à de bonnes corrélations entre les item.

C'est de cette dernière conception de la validité de contenu qu'origine la définition donnée par l'Association Américaine de Psychologie (3). "La validité de contenu est obtenue au moyen d'une description de l'univers des item à partir duquel un choix fut effectué, incluant une description du processus et des critères de choix". Cette définition met donc en lumière l'importance de définir l'univers des questions à mesurer d'une part et la nécessité de décrire le processus et les critères ayant conduit à l'acceptation ou au rejet des item composant le test final. C'est d'ailleurs à partir de cette définition que Lennon (34) définit trois postulats sous-tendant l'utilisation du concept de validité de contenu.

Postulat 1: le domaine auquel le constructeur de test est intéressé doit être traduisible sous forme d'un univers de réponses définissable et significatif.

Postulat 2: un échantillon peut être tiré de cet univers d'une façon compréhensible et dirigée (purposive).

Postulat 3: l'échantillon et le processus d'échantillonnage peuvent être définis avec précision de façon à ce que l'utilisateur puisse juger jusqu'à quel point les réponses à l'échantillon sont représentatives des réponses à l'univers des item.

Nous approfondirons ultérieurement ces postulats. Pour l'instant, c'est l'évolution du concept de validité de contenu qu'il est important de saisir. C'est ce dernier qui est actuellement accepté. Cette option accorde une grande importance à la pertinence du contenu et par le fait même, donne

moins d'importance aux approches purement statistiques; c'est ce qui fait dire à Nunnally (42) que la validité de contenu est habituellement établie par des moyens autres que la statistique.

Parallèlement à ce type de validité de contenu, s'insère le concept de validité de face. Anastasi (4) explicite ce concept de façon claire en avançant le caractère complémentaire de ce type de validité. Ce concept a pour objet le caractère apparent d'un test; il veut répondre à la question: est-ce que les item d'un test apparaissent pertinents aux personnes qui auront à utiliser le test?

En conclusion, il est bon de noter la distinction apportée par Goodenough (27) entre les concepts de validité intrinsèque et extrinsèque. La validité intrinsèque appartient à un test où "l'expérimentateur est intéressé par le type de comportement manifesté dans la performance au test"; dans ur tel cas, le critère est interne au test, d'où l'adjectif "intrinsèque" et la validité du test a rapport au degré de représentativité des item inclus dans ce test.

On parlera de validité extrinsèque lorsque l'expérimentateur ne s'intéresse à un type de performance qu'en autant que ce dernier est relié à quelque critère externe, tel l'intelligence; la validité prédictive et concomittante sera alors d'intérêt premier alors que la validité de contenu s'applique au caractère "intrinsèque" d'un test.

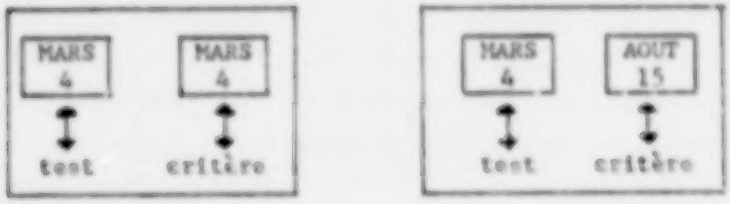
Jetons maintenant un coup d'oeil sur la validité prédictive et concomittante.

II - Validité prédictive et concomittante

Ces deux types de validité s'appliquent aux tests auxquels Goodenough (27) applique le concept de validité extrinsèque. Ils concernent les relations existant entre des résultats à un test et des critères extérieurs qu'on veut prédire par ces résultats. Campbell (9) parle alors de validité

pratique parce que les variables extérieures, servant de critères, sont habituellement très pratiques (succès scolaire, compétence, etc....).

La différence entre la validité prédictive et concomittante est une différence temporelle. Le graphique suivant illustre bien cette différence.



L'élaboration de ces deux concepts n'a pas suivi le long cheminement de celui de validité de contenu. Très tôt, le caractère particulier de ces deux types de validité a été reconnu; c'est au niveau des techniques d'analyse et du choix des critères que les efforts de recherche sont concentrés. Ce n'est pas l'objet du présent document de travail d'approfondir ces deux questions. Il serait quand même utile d'introduire le concept de validité synthétique introduite par Lawshe (32) et explicité par Balma (5). Ce type de validité peut être utilisé lorsque des résultats sont mis en relation avec un critère complexe incluant plusieurs éléments. Les techniques associées à ce type de validité permettent de synthétiser en un seul coefficient un ensemble de coefficients concomittants ou prédictifs. Le lecteur intéressé trouvera dans Lawshe et Balma (33) une description de ces techniques.

En conclusion, il est important de noter le lien étroit existant entre ces deux types de validité et la validité de contenu. Ce lien est d'ordre temporel; en effet, effectuer une étude de prédiction suppose qu'a déjà été vérifiée la validité de contenu à la fois du test lui-même et du critère utilisé. On ne peut donc ignorer le contenu en faveur d'une étude prédictive.

III - Validité de construit

Ce type de validité est le plus récent des concepts développés par l'APA (3); il est aussi très controversé. Essentiellement, il a un rapport étroit avec le trait théorique qu'un test veut mesurer. En ce sens, c'est un développement de la validité logique de contenu, telle qu'explicitée par Lennon (34) et Lord (37). Ce type de validité a pour but d'établir une relation entre la performance de sujets à un test et l'interprétation de cette performance en termes de variables théoriques. Ces variables théoriques peuvent être très variées, mentionnons-en quelques-unes, telles l'anxiété, la dépendance, l'intelligence, l'insatisfaction, l'acuité perceptuelle, etc....

Le terme "validité de construit" a été introduit par l'APA (3) en 1954. Ce comité indique que "la validité de construit est habituellement étudiée pour augmenter la compréhension des qualités psychologiques mesurées par le test. Son étude devient pertinente lorsqu'il n'existe pas de mesure-critère de ces qualités en ce que ces mesures-critères sont nombreuses" (2). Cronbach et Meehl (15) explicite autrement ce concept; pour eux, il faut se préoccuper de validité de construit lorsque des résultats à un test sont interprétés comme la mesure d'un trait ou d'une qualité non définie opérationnellement. Selon eux, un construit doit posséder au moins les trois caractéristiques suivantes: premièrement, le construit est un attribut postulé qui doit être réflété dans la performance au test; deuxièmement, un construit a des propriétés prédictives et troisièmement, la signification d'un construit est donnée à partir des lois qui sont à sa base.

Bechtoldt (7) prend un malin plaisir à démontrer le caractère confus de ces définitions et des propriétés. En effet, la relation entre la performance à un test et un attribut postulé nous éclaire peu sur la nature de cet attribut, mais nous renseigne sur une définition particulière de cet attribut; pour un seul attribut, il existe plusieurs théories, plusieurs modèles dont les expérimentations respectives tendent à en prouver la valeur; habituellement, c'est que ces différents expérimentateurs accolent à un même attribut des définitions différentes. Finalement, comme le souligne Lord (36), toute relation entre un attribut et des résultats à un test ne sert qu'à éta-

blir une définition opérationnelle de cet attribut. Bechtoldt (7) critique de la même façon les propriétés prédictives et les autres concepts avancés par l'APA (3) et Cronbach et Meehl (15). Ce type de critique, tout en étant bien appuyé, ne justifie pas pour autant sa position extrême de rejeter la formulation du concept de validité de construit. Il faut plutôt voir dans ces imprécisions la complexité des différents phénomènes étudiés en psychologie. Le concept de validité de construit présente une façon d'encourager les expérimentateurs à approfondir la signification des mesures (Ebel (19)) qu'ils effectuent de façon à faciliter leur interprétation.

Si l'on accepte ce point de vue, le concept de validité de construit acquiert une dimension dynamique et longitudinale dont les éléments sont les suivants: un modèle théorique**, des mesures empiriques, des relations entre des variables extérieures et le modèle ou les mesures effectuées.

C'est dans cette optique que Campbell et Fiske (10) ont défini leurs concepts de validité convergente et discriminante et proposé une méthodologie permettant d'établir ces deux types de validité. Ces deux types de validation mettent en relation les trois éléments de validité de construit, cités précédemment; ils permettront de vérifier que

a) les résultats sont en corrélation avec des variables empiriques qui, d'après le modèle utilisé, sont reliées à l'attribut mesuré;

b) les résultats sont en corrélation avec d'autres résultats selon des hypothèses du modèle (validité convergente);

c) les résultats ne sont pas en corrélation avec d'autres résultats selon les hypothèses du modèle (validité discriminante).

En conclusion à cette partie théorique portant sur le concept de validité, il est intéressant de présenter une reformulation des dimensions de ce concept, proposée par Cattell (11). Cette reformulation permettra de clarifier trois dimensions du concept de validité.

** Le lecteur intéressé pourra consulter l'intéressant article de Simon et Newell (46) portant sur les différents types de modèles.

La première dimension touche le caractère concret (particulier) ou abstrait (conceptuel) d'un test. On reconnaîtra la parenté entre cette validité abstraite et le concept de validité de construit. Mesurer l'intelligence a un caractère abstrait, mesurer le bégaiement a un caractère concret.

La deuxième dimension touche le caractère naturel ou artificiel d'un test ou d'un critère. L'auteur veut ici rappeler que le but premier de la psychologie est la prédiction de comportements réels, il souhaite donc que les chercheurs mettent un accent prononcé sur le choix de critères et de tests naturels. Il critique, en ce sens, les études de validation portant sur des tests d'intelligence dont la validité est assumée à partir de tests d'intelligence périmés. Cependant, il faut noter que le caractère naturel ne va pas à l'encontre du caractère abstrait d'un test ou d'un critère.

Enfin, la troisième dimension touche le caractère direct ou circonstantiel (indirect) d'un test ou d'un critère. Une étude de validation aura toujours avantage à utiliser des critères indirects; un exemple est apporté par Lord (37) lorsqu'il suggère de mettre un test en corrélation avec des variables qui, selon le modèle utilisé, ne doivent pas être reliées au test étudié.

Ces trois dimensions nous invitent à prêter une attention particulière aux critères utilisés dans une étude de validation ayant pour objet d'approfondir nos connaissances à propos d'une mesure effectuée et d'expliciter par cette mesure un aspect réel d'un comportement.

PARTIE 2

Validation du test PERPE

Introduction

La première partie de ce texte a présenté diverses conceptions du problème de la validité, le cheminement qui y est suivi peut être considéré comme complet. Cette deuxième partie n'aspire pas au même objectif; elle prendra réellement l'allure d'un document de travail, isolant d'une part des problèmes précis, amorçant d'autre part des solutions possibles à ces problèmes: solutions qui devront être discutées, puis vérifiées empiriquement. Le ton de cette deuxième partie pourra apparaître critique, mais son but sera toujours de préciser des problèmes et d'amorcer des solutions.

Elle se compose de six parties. La première apporte une description du test PERPE collégial, la deuxième abordera le problème des objectifs visés par l'utilisation de ce test, la troisième traitera de la validité de contenu du test, la quatrième, de la validité prédictive et concomittante, la cinquième, de la validité de construit, la sixième présentera des éléments de conclusion.

Les différents types de validité qui seront étudiés le seront dans l'optique décrite dans la première partie de ce texte.

I - Description du test

item

Le test 'PERPE' est composé de 61 item. Ces 61 item sont habituellement qualifiés de dimensions dans les diverses publications effectuées par les membres de ce groupe de recherche.

Chaque item est composé de cinq éléments: un titre, identifiant l'item; une description, visant à clairifier le

sens de ce titre; une description des cinq points composant l'échelle d'évaluation; deux questions, la première demandant à l'étudiant de situer le cours (ou le professeur) sur l'échelle d'évaluation, la deuxième demandant où il devrait se situer pour qu'il (étudiant) soit satisfait.

sections

Les 61 item ont été regroupés en six sections: encadrement (12 item), versatilité (12 item), magnétisme (10 item), altérocentrisme (12 item), climat (10 item) et disponibilité (5 item).

mesures

Plusieurs mesures sont effectuées sur les réponses des étudiants au niveau de chaque item, au niveau des sections et au niveau de tout le test.
Au niveau de chaque item, huit mesures sont transmises à l'utilisateur:

Exemple:

Titre	item	MA	MB	%	MAB	MAL	R	G	S
Quantité de matière	01	3.18	3.36	36	0.36	-0.18		7	7

Dans cet exemple fictif, nous avons une illustration de ces huit mesures.

1) MA: moyenne des réponses des étudiants à la question 1 de chaque item.

2) MB: moyenne des réponses des étudiants à la question 2 de chaque item.

3) % : proportion d'étudiants dont la réponse à la question 1 diffère de la réponse à la question 2.

4) MAB: moyenne des écarts en valeur absolue de tous les étudiants entre leurs réponses à la question 1 et leurs réponses à la question 2.

5) MAL: moyenne des écarts en valeur relative de tous les étudiants entre leurs réponses à la question 1 et leurs réponses à la question 2.

6) R : cette colonne peut contenir trois informations, un blanc, un R (rejets) ou un C (contradiction). Un C apparaît dans cette colonne lorsque, parmi les étudiants dont la réponse à la question 1 diffère de celle à la question 2, la proportion d'étudiants dont la différence entre les deux questions est positive ou négative est supérieure ou égale à 33% et inférieure ou égale à 67% ($33\% \leq p \leq 67\%$).
Un R apparaît dans cette colonne lorsque 25% ou plus des étudiants n'ont pas répondu à l'item analysé. Il est nécessaire de noter que, si un item remplit ces deux conditions (trop de rejets et de contradiction), c'est l'indice R qui apparaît seul, étant donné que ce R indique que l'ensemble des résultats sur cette dimension sont peu valides.

7) G : cote-stanine de l'utilisateur pour l'item analysé comparant la MAL qu'il obtient à ceux de tous les professeurs ayant utilisé le test.

8) S : cote-stanine comparant la MAL d'un professeur à celles de tous les professeurs enseignant dans la même matière que lui et qui ont utilisé le test.

Au niveau de chacune des six sections, six mesures sont effectuées:

Exemple

Sections	Sect.	%	MAR	MAL	G	S
Encadrement	I	32	0.39	0.29	6	6
Versatilité	II	43	0.55	0.46	6	6
Magnétisme	III	33	0.41	0.40	6	6
Altérocentrisme	IV	27	0.32	0.32	7	7
Climat	V	42	0.46	0.31	5	5
Disponibilité	VI	33	0.47	0.47	7	7

% = cette mesure consiste à faire la moyenne des pourcentages des dimensions de chaque section en excluant celles marquées d'un R.

MAB, MAL = ces deux mesures sont obtenues de la même façon en calculant la moyenne des MAB et MAL des diverses dimensions de chaque section, excluant celles marquées d'un R (cependant, pour la MAL, la moyenne est calculée à partir des valeurs absolues des MAL des item).

G, S = ces deux cotes-stanines mettent en comparaison les MAL de sections de chaque utilisateur et celles de tous les autres utilisateurs du test (en général ou par matière, selon le cas). Les MAL de sections exclues de ce calcul sont celles dotées d'un R.

R = un R indique le rejet d'une section. Cette dernière est rejetée lorsque 33% et plus des dimensions utilisées pour calculer la MAL de section sont rejetées.

Au niveau de l'ensemble du test, cinq mesures sont effectuées:

% = ce pourcentage indique la moyenne des pourcentages de toutes les dimensions non-dotées d'un R pour chaque dossier.

MAB, MAL = ces deux mesures sont obtenues de la même façon, à partir des MAB et MAL de chaque dimension non dotée d'un R. Pour la MAL, la moyenne est calculée à partir des valeurs absolues.

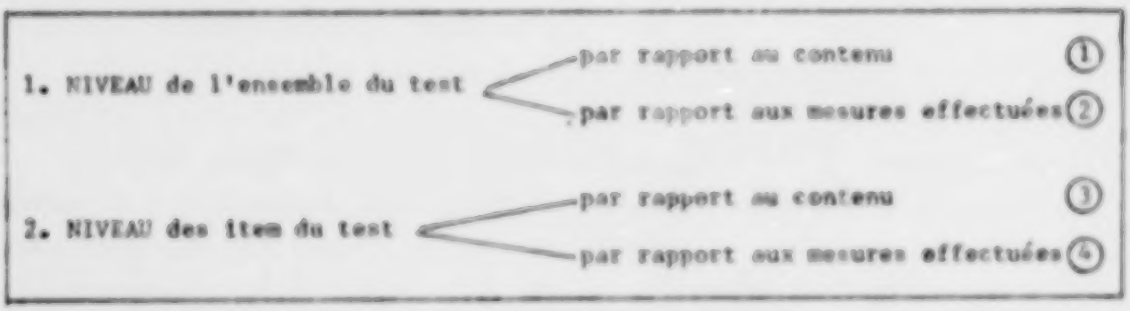
G, S = ces deux cotes-stanines mettent en comparaison les MAL générales de chaque utilisateur et celles de tous les autres utilisateurs du test (en général ou par matière, selon le cas).

Plusieurs autres éléments pourraient être inclus dans une description du test. Ces informations intéressantes dépassent pour l'instant l'objet de ce document qui visait une description opérationnelle de la structure du test et des mesures effectuées. Nous allons maintenant faire un pas supplémentaire et clarifier les objectifs de ce test.

II - Les objectifs du test PERPE

Le cheminement suivi pour clarifier ces objectifs consiste à identifier comment ces objectifs ont été formulés dans les diverses publications des membres de PERPE. Une analyse de ces formulations, appuyée d'une réflexion sur le contenu du test nous aidera ensuite à proposer une terminologie dont les publications subséquentes devraient s'inspirer.

Ces objectifs sont cités dans les diverses publications soit à l'occasion de description du test pris dans son ensemble, soit à l'occasion de description des item du test, soit à l'occasion de l'explication des mesures effectuées par le test. Il faut donc distinguer ces trois niveaux. Le tableau suivant résume ce plan.



Les notes composant le tableau suivant, tirées des diverses publications, rapportent les formulations d'objectifs selon les niveaux décrits dans le plan ci-dessus.

1. Ensemble du test (contenu)

- "notre principal souci fut de voir à ce que les dimensions représentent adéquatement la relation professeur-étudiant" (13, p. 167)
- "mesurer la relation professeur-étudiants, ou... la situation d'enseignement" (12, p. 3)
- "renseigner le professeur sur les perceptions qu'ont les étudiants à propos de la réalité du cours et à propos de leur idéal face au cours"(12, p. 3)

- "indiquer... au professeur comment ses étudiants le perçoivent, quelles sont leurs attentes et sur quoi porte leur insatisfaction"(26, p. 5)
- "construire un test d'évaluation des cours"(25, p. 116)
- "... test d'évaluation de la relation professeur-étudiant" (25, p. 116)
- "... évaluer la satisfaction des étudiants à l'égard de la relation professeur-étudiant (20, p. 71)
- "questionnaire d'évaluation des cours" (23, p. 3)
- "dimensions évaluent la relation professeur-étudiant" (23, p. 3)
- "... instrument mesurant les perceptions étudiantes de la relation professeur-étudiant"(44, p. 10)
- "...certaines composantes de la situation éducative" (44, p. 10)

Comme il est facile de l'observer, on parle indifféremment d'évaluation du professeur, du cours, de la relation professeur-étudiants, de la situation d'enseignement, de la situation éducative. Ces termes sont-ils tous synonymes? Préciser l'objectif par les fréquences d'apparition de ces différents termes nous conduirait à opter pour la relation professeur-étudiant. C'est une solution trop facile. Deux solutions plus adéquates s'offrent pour résoudre ce problème; la première consiste à choisir l'une des terminologies utilisées, à la préciser et à la confronter au contenu même du test; la deuxième consiste à effectuer une analyse de contenu du test, à en identifier les éléments qui serviraient à définir opérationnellement un terme qui engloberait ce contenu. Il est actuellement trop tôt pour choisir l'une ou l'autre méthode; l'analyse de la validité de contenu pourra par contre nous éclairer quelque peu. Cette analyse fera l'objet de la partie suivante.

2. Ensemble du test (mesures effectuées)

- "les mesures de réalité, d'attentes et d'insatisfaction constituent la base des informations..." (13, p. 164)
- "pour mesurer le degré de satisfaction d'un étudiant, ... l'écart

entre son évaluation de la réalité et celle de ses attentes" (13, p. 164)

- "l'insatisfaction est donc construit... en fonction d'une théorie quelconque qui justifie de telles inductions (12, p. 3)
- "les scores A et B caractérisent la perception des étudiants quant à la réalité d'un cours et quant à leurs attentes "(12, p. 61)
- "considérer la moyenne d'un groupe (MA) comme représentative de la réalité... repose" (12, p. 61)
- "le construit fondamental... est celui du continuum satisfaction-insatisfaction" (12, p.61)
- "rationnel du score de satisfaction..." (12, p. 60)
- "le score d'insatisfaction correspond..." (12, p. 63)
- "évaluent comment ils perçoivent le comportement... quelles sont leurs exigences ou attentes... dans quelle mesure ils sont satisfaits" (24, p. 66)
- "mesure le "degré de satisfaction"" (25, p. 119 et 124)
- "professeur reçoit une cote globale de satisfaction" (25, p. 119)
- "mesure l'insatisfaction des étudiants" (25, p. 134)
- "mesures indépendantes de réalité et d'attentes (25, p. 134)
- "cote générale de satisfaction" (25, p. 145)
- "évaluer la satisfaction des étudiants à l'égard de la relation..." (20, p. 71)
- "mesure le degré de satisfaction" (23, p. 13)
- "mesure les perceptions étudiantes de la réalité et leurs attentes" (23, p. 13)

Quoique ambigue, la terminologie utilisée ici nous éclaire bien sur les concepts devant faire l'objet d'une validation de construit. En effet, il faudra d'abord préciser "la théorie qui justifie de telles inductions"; par la suite, préciser le vocabulaire utilisé. Au niveau des mesures A, prises individuellement, doit-on parler de mesure de réalité ou de perception des

attentes? Le cumul des A et B en MA et MB permet-il l'interprétation en termes de réalité, d'attentes? La mesure A-B est-elle une mesure de satisfaction ou d'insatisfaction; les deux termes sont-ils équivalents? avons-nous un continuum satisfaction-insatisfaction? Les mesures A et B sont-elles indépendantes? Autant de questions à propos duquel l'utilisateur du test ne peut répondre actuellement devant l'imprécision et la multiplicité des termes employés.

3. Item du test (contenu)

- "chaque dimension rejoint un aspect différent de..."(13, p. 162)
- "il fallait d'abord connaître les aspects..." (26, p. 5)
- "le test analyse 61 dimensions de l'enseignement" (24, p. 66)
- "61 aspects différents..." (25, p. 119)
- "les questions évaluent 61 aspects (23, p. 13)
- "mesure certaines composantes de..." (44, p. 10)

En fait, seuls 2 termes sont employés: on parle soit de dimensions, soit d'aspects ou de composantes de la situation d'enseignement. Je voudrais attirer l'attention ici sur l'usage inopportun du terme "dimension". Si on se fie au Larousse, le terme 'dimension' réfère à un élément nécessaire pour évaluer quelque chose; ainsi, la hauteur d'un triangle est un élément nécessaire pour évaluer sa surface. Ce n'est que lorsque les 61 item du test auront été jugés nécessaires à l'évaluation de la mesure effectuée par l'ensemble du test que ceux-ci pourront être qualifiés de dimensions; d'ailleurs, le terme "dimension" est souvent identifié à celui de "facteur". Je propose donc d'utiliser le terme "aspect" ou "composante" pour parler de chaque item du test.

4. Item du test (mesures effectuées)

Les références se rapportant à cette partie sont sensiblement les mêmes que celles de la partie 2. Les questions soulevées sont aussi sensiblement les mêmes; une réponse A ou B doit-elle être interprétée en termes de réalité ou d'attentes ou bien en termes de perceptions. La mesure A-B

exprime-t-elle un degré de satisfaction ou d'insatisfaction?

Cette brève incursion de la terminologie utilisée ne nous a pas permis de préciser les objectifs du test PERPE; par contre, elle a permis de clarifier des questions auxquelles devront répondre nos analyses des divers types de validité. Arrêtons-nous maintenant sur ces problèmes.

III - La validité de contenu

Nous avons discuté précédemment le concept de validité de contenu. Cette discussion a permis d'isoler trois postulats; le premier affirme qu'il est possible de définir l'univers des item, du moins ses caractéristiques, qui est à l'origine de ce contenu. Le deuxième définit la méthodologie sous-jacente au choix des item qui composeront l'échantillon retenu. Le troisième demande de décrire avec précision le processus d'échantillonnage utilisé. Ces trois postulats feront l'objet de deux parties. Une troisième partie parlera de la validité de face de notre instrument. Une quatrième identifiera les liens entre ce type de validité et le construit à mesurer à partir des notions de validité intrinsèque et extrinsèque.

A) Univers des item

Il est bien difficile de parler d'un univers d'item; en effet, comment espérer établir une liste de tous les item mesurant un attribut quelconque (v.g. intelligence, situation de cours,...). C'est pourquoi les constructeurs de tests définissent cet attribut en décrivant ces principales caractéristiques. L'utilisateur pourra alors s'inspirer de ces caractéristiques pour connaître ce que le test veut mesurer. Dans notre cas, l'utilisation de plusieurs termes (relation professeur-étudiant, situation de cours, situation d'enseignement...) se rapportant à un seul contenu nous invite à préciser les caractéristiques de ce contenu; les courtes notes historiques qui suivent ont cet objectif.

notes historiques

Les deux sources principales servant à ces notes sont les documents de Chabot (12) et Gagnon (26).

C'est à l'intérieur d'un cours de psychologie sociale qu'est née chez l'auteur du test l'idée d'un instrument d'évaluation des professeurs par les étudiants. Plus précisément, c'est au cours de discussions en équipes portant sur "les critères sur lesquels se basent les étudiants pour évaluer un professeur" (Gagnon (26), p. 22). Afin d'identifier les divers éléments à mesurer, l'auteur, selon Gagnon, précise d'une part la signification de ces éléments et décide de consulter prioritairement les premiers observateurs: les étudiants. Par éléments, l'auteur vise "des aspects précis du cours ou de la personnalité du professeur qui influencent la perception des étudiants". Cette définition permet d'opérationnaliser le terme plus large, plus global habituellement utilisé. Au sujet de ce terme, il serait souhaitable de l'uniformiser afin d'éviter l'ambiguité actuelle dûe à la multiplicité des termes utilisés. Pour ma part, je parlerais de situation d'enseignement. Ce terme m'apparaît moins restrictif que "situation de cours" ou "relation professeur-étudiants; en effet, le terme "situation de cours" semble référer à des aspects uniquement relié à un enseignement traditionnel alors que plusieurs item pourrait s'appliquer à d'autres types d'enseignement. Pour sa part, le terme "relation professeur-étudiants" semble mettre la priorité sur des aspects d'interaction entre ces deux secteurs alors que les item couvrent beaucoup d'aspects extérieurs à cette interaction.

Le contenu du test mesure donc la situation d'enseignement, soit des aspects précis du cours ou de la personnalité du professeur qui influencent la perception des étudiants. Voilà donc défini notre univers des item. Evidemment, il serait idéal de situer cette définition dans un modèle de la situation d'enseignement; ce travail, énorme en soi, serait quand même facilité par l'approfondissement des modèles décrits par Gage (22). Une lecture rapide de ce chapitre nous permet quand même de voir que notre définition de l'univers du contenu tient compte d'aspects importants de la situation d'enseignement, telle que définie dans ces modèles.

B) Méthodologie et processus d'échantillonnage

Maintenant que nous avons réussi à identifier l'univers du contenu et ses principales caractéristiques, nous allons étudier la méthodologie utilisée pour obtenir les item composant le test.

L'auteur a consulté les étudiants de son cours de psychologie sociale pour recueillir des informations sur les aspects qui influencent leurs perceptions de leurs cours ou de la personnalité de leurs professeurs. Une fiche de notation leur fut remise avec la consigne d'y noter des faits de comportement qu'ils jugent importants et qui sont "applicables à tout cours et à tout professeur" (12). Cette fiche comportait une description du comportement et une classification (contenu, interaction, personnalité). Environ 600 fiches furent ainsi recueillies.

Ces 600 fiches furent analysées, les contenus équivalents furent éliminés, ceux qui étaient trop généraux furent précisés; le bilan produisit 53 item différents. Huit item furent ajoutés par l'équipe de rédaction composée de l'auteur et d'étudiants parce qu'ils furent jugés importants pour évaluer la situation d'enseignement. Ces 61 item furent expérimentés dans la forme A du test. Le tableau suivant présente les regroupements d'item utilisés tout au long des six administrations du test.

forme	N. catégorie	n. item	Titres
A	4	20	Caractéristiques visant la personnalité du prof
		13	Caractéristiques visant le contenu du cours
		16	Caractéristiques visant la présentation du contenu
		12	Caractéristiques visant l'interaction prof-élèves
B	4	20	Le titre n'apparait pas dans le quest.(Gagnon,
		13	20, p. 57). Les regroupements sont les mêmes que
		16	pour la forme A. Cette politique est appliquée
		12	à toutes les formes ultérieures.
C	6	11	Emetteur personnel
		10	Emetteur pédagogique
		10	Message - contenu
		14	Message - forme technique
		8	Message - forme interpersonnelle
		8	Relations interpersonnelles (d'ordre socio-émotif)
D	6	12	Mêmes regroupements que la forme C
		11	
		10	
		13	
		8	
		7	
F	6	9	Message: ce qu'il est
		11	Message: comment il est transmis
		19	Emetteur: ce qu'il est
		10	Emetteur: comment il transmet
		14	Interaction: dimension cognitive
		8	Interaction dimension socio-émotive

Forme	N. catégorie	N. item	Titres
G	6	12	Encadrement
		12	Versatilité
		10	Magnétisme
		12	Altérocentrisme
		10	Climat
		5	Disponibilité

Les regroupements des cinq premières formes s'appuient surtout sur une base logique; ceux de la forme G s'inspire surtout des intercorrélations entre les item. Il est intéressant de noter que les regroupements de la forme G apparaissent beaucoup plus interreliés (surtout les 4 derniers regroupements) que ceux des formes précédentes. Malgré tout, on devrait s'attendre à une plus grande homogénéité des regroupements de la forme G; c'est d'ailleurs sur cette base d'homogénéité qu'il est justifié de calculer des scores de section.

Chacune des six formes du test comprend 61 item. Si on considère ceux de la forme G, on remarque que 37 d'entre eux originent, avec des modifications mineures, de la forme A du test, et que 24 ont été ajoutés tout au long des expérimentations successives. Ces 24 item proviennent soit de suggestions de professeurs, d'examen de la littérature (Ryans (45), Solomon et al. (48)) ou de suggestions issues de l'équipe de rédaction.

L'élimination d'item d'une expérimentation à l'autre n'est donc pas basée sur le contenu de ces item; cette élimination est reliée à plusieurs critères psychométriques définis dans Gagnon (26), p. 11. Evidemment, il est probable que la signification des critères utilisés ait un lien avec le contenu de certains item (v.g. certain contenu comme "propreté" ne présente que très peu d'insatisfaction), cependant l'interprétation de ces contenus se situera dans la non-influence de ces contenus envers l'évaluation que des étudiants portent sur des professeurs.

Nous pouvons maintenant étudier plus précisément la notion de modèles, reliés à la validité de contenu.

Deux cheminements sont possibles en vue de l'élaboration d'un modèle relié à la validité de contenu.

Le premier consiste à étudier les principales caractéristiques explicitant l'univers de conteu; cette étude a pour but d'approfondir la signification de ces caractéristiques en les reliant à des modèles déjà existants. Rappelons que les trois caractéristiques de notre univers de contenu sont la personnalité du professeur, le contenu et l'interaction. Ces dénominations se sont complétées peu à peu lors de l'élaboration des formes C-D et F du test.

Le deuxième cheminement, plus empirique, part de résultats recueillis pour identifier des "patterns" représentatifs du contenu des item du test. C'est à partir de techniques comme le sociogramme, le "cluster analysis", l'analyse factorielle que cette démarche est concrétisée. C'est d'ailleurs cette approche qui a été favorisée lors de l'élaboration des six regroupements de la forme F du test.

Dans la réalité, l'élaboration de notre modèle doit s'inspirer de ces deux cheminements. En effet, ne s'inspirer que du premier enleverait à notre modèle toute la rigueur d'une vérification empirique; ne s'inspirer que du deuxième n'ajouterait à notre compréhension, qu'une formulation statistique. C'est pourquoi nous proposons la démarche suivante:

1 - Comparaison des trois catégories et des six sous-catégories (forme F) à des catégories utilisées par d'autres chercheurs ayant aussi pour but la construction d'un instrument d'évaluation. On peut trouver de telles utilisations dans Hamachek (29), Medley et Mitzel (40), Ministère de l'éducation (41), Fox et Brookshire (21), Barr et al.(6). Cette comparaison permettrait d'une part de préciser la signification des catégories utilisées par PERPE et d'autre part d'initier l'identification de catégories plus représentatives de notre contenu et de l'interaction entre ces catégories.

2 - Elaboration parallèle de deux ou trois modèles du genre de celui proposé par Lamontagne (30). On pourrait s'inspirer aussi de ceux de Blake et Mouton (8), de Runkel (voir Gage (22)), de Costin (14). Ce type de recherche permettrait d'offrir des éléments de compréhension parallèles et d'élargir les possibilités d'arriver à la meilleure compréhension de notre contenu.

3 - Cette étape en est une de vérification empirique des modèles proposés précédemment. Elle implique le choix d'une technique appropriée et son utilisation sur les deux ou trois modèles élaborés.

Cette démarche permettrait à mon avis d'éviter les écarts entre notre catégorisation logique de la forme F et notre catégorisation empirique de la forme G.

C) Validité de face

Chabot (12), sous l'en-tête "appréciation du contenu par les professeurs" donne les résultats d'un questionnaire-sondage administré à un échantillon restreint de professeurs. L'analyse de ces données fournit un indice précieux, confirmant ce type de validité. Et même si ces questions ne portaient pas comme telle sur le contenu du test, les résultats justifient cette validité.

D) Conclusion

En analysant le contenu du test, nous sommes préoccupés de la signification des item, des comportements reliés à ces contenus, des regroupements de divers item en des catégories; ce type d'analyse se réfère à la validité intrinsèque, proposée par Goodenough.

Par contre, des mesures sont effectuées sur ces contenus, des perceptions de réalité, d'attentes et une mesure d'insatisfaction. Au cours des diverses expérimentations, des item furent éliminés pour des raisons d'ordre psychométrique. Les réponses à A et B devenaient des signes d'insatisfaction. Ce type d'analyse se réfère à la validité extrinsèque, proposée par Goodenough.

Nous approfondirons plus loin (validité de construit) la signification des mesures effectuées, mais il faut retenir que le contenu et sa validité sont influencées par celle des mesures effectuées.

IV - Validité prédictive et concomittante

Ces types de validité sont habituellement associés à des objectifs utilitaires. Se rappelant nos discussions sur les différences entre ces

deux types de validité, nous pouvons les discuter simultanément. En effet, prédire la compétence appartiendra à la validité concomittante si le critère de compétence est simultané aux mesures recueillies par PERPE et à la validité prédictive si le critère se situe à une période ultérieure.

Ces deux types de validité s'emploient surtout dans l'industrie ou dans des programmes de formation en éducation. On cherche habituellement à prédire le succès scolaire, la réussite au travail. Dans un type de recherche comme la nôtre, les variables extérieures sont plutôt utilisées pour approfondir notre compréhension de la situation éducative ou des mesures que l'on utilise. C'est pourquoi des variables, qui pourraient être utilisées dans une optique prédictive, sont utilisées ici dans une optique d'approfondissement du lien entre ces variables et les mesures effectuées.

Les variables étudiées peuvent être très variées, celles du PAS en sont un exemple. On peut en ajouter d'autres comme la "compétence" du professeur, bien discutée dans Chabot (12), la classification des professeurs en "good" et "bad"... etc. Cependant, de telles recherches, si elles nous éclairent sur les mesures de PERPE, restent loin d'objectifs purement prédictifs.

V - Validité de construit

Cette section se divise en cinq parties. La première définira une terminologie des termes utilisés dans cette section, la deuxième abordera les modèles généraux de la situation éducative, la troisième étudiera la validité des mesures d'insatisfaction, la quatrième abordera les mesures de réalité et d'attentes, la dernière partie apportera une conclusion à cette section.

A) Terminologie

Cette partie veut répondre, en ce qui concerne les mesures effectuées dans le test, à l'ambiguité des termes, telle qu'illustrée dans la partie II. Les différents aspects mesurés par le test permettent d'obtenir des mesures au niveau de chaque étudiant, au niveau d'une classe d'étudiants,

au niveau de chaque aspect, au niveau de certains regroupements d'aspects et enfin, au niveau de l'ensemble du test.

Au niveau desmesures A et B, le vocabulaire utilisé se limite à "perception de la réalité" et "perception des attentes"; ces termes sont justes puisqu'ils ne font qu'exprimer opérationnellement la consigne du test. Par contre, au niveau des mesures de classes, on parle parfois de perceptions moyennes, parfois de mesures de la réalité, se basant sur Chabot (12, p. 61). Nous croyons prématuré de conclure qu'une moyenne de perceptions puisse être représentative de la réalité sur de simples considérations logiques: l'adoption d'un tel postulat nous semble peu rigoureux tant que ces mesures n'ont pas été confrontées à d'autres mesures de réalité. Pour l'instant, il faut donc parler de moyenne de perceptions de la réalité et d'attentes.

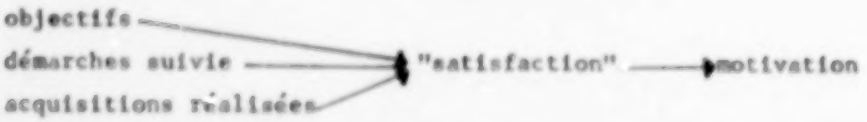
Quant au construit A-B, il ne semble y avoir aucune consistence des termes utilisés. A notre avis, il est absurde de parler de continuum satisfaction-insatisfaction (12, p. 60), de degré de satisfaction ou de score de satisfaction (13, p. 6), (12, p. 60), (24, p. 66), (25, pp. 119-124-145), (20, p. 71), (23, P. 13). En effet, le construit A-B interprété en termes de différences entre A et B, nous oblige à parler d'insatisfaction lorsque cette différence est $\neq 0$ et de satisfaction lorsque cette différence est égale à 0. Nous n'obtenons donc jamais de degré de satisfaction sauf peut-être par inférence en disant que lorsque A=B=1 la satisfaction est plus petite que lorsque A=B=4; un tel raisonnement, non appuyé sur une base théorique, est du domaine de la spéculation et doit être abandonné. Nous parlerons donc de degré d'insatisfaction et associerons le construit A-B à une mesure d'insatisfaction.

B) Modèles de la situation éducative

La validité de construit peut se situer à deux niveaux. Le premier situera l'insatisfaction étudiante dans la situation d'enseignement. Le deuxième étudiera plus spécialement les mesures utilisées par PERPE et les contenus mesurés. Cette partie traitera du premier niveau.

Dans les écrtis de PERPE, deux documents (Lamontagne (31) et Foucher (20)) esquissent une approche de ces modèles.

Foucher (20), pour sa part, étudie la "satisfaction" (hic!) comme une composante de l'apprentissage. Ce texte, sous forme de réflexions, nous permet quand même de schématiser le modèle suivant



L'auteur définit l'apprentissage comme un processus permettant une modification de l'individu. Ce processus influence la "satisfaction" de l'étudiant, laquelle "satisfaction" influence sa motivation. La brièveté de l'article nous empêche cependant de comprendre plus avantageusement cette conceptualisation; ce modèle reste incomplet et imprécis: il faut noter cependant que l'objet de l'article n'était pas d'élaborer un tel modèle; cependant, les éléments qu'il mentionne aurait avantage à être développés.

Lamontagne (31), pour sa part, offre une conceptualisation plus théorique; il suggère 4 modèles dont les éléments sont autant des mesures subjectives du test que des paramètres behavioraux. Cependant, ces quatre modèles représentent quatre possibilités parmi lesquelles il faudra faire un choix. Ce n'est pas l'objet du présent document d'effectuer ce choix à la lumière de données empiriques. Par contre, la méthodologie proposée dans les prochaines parties pourra déjà nous éclairer en rapport à ce choix.

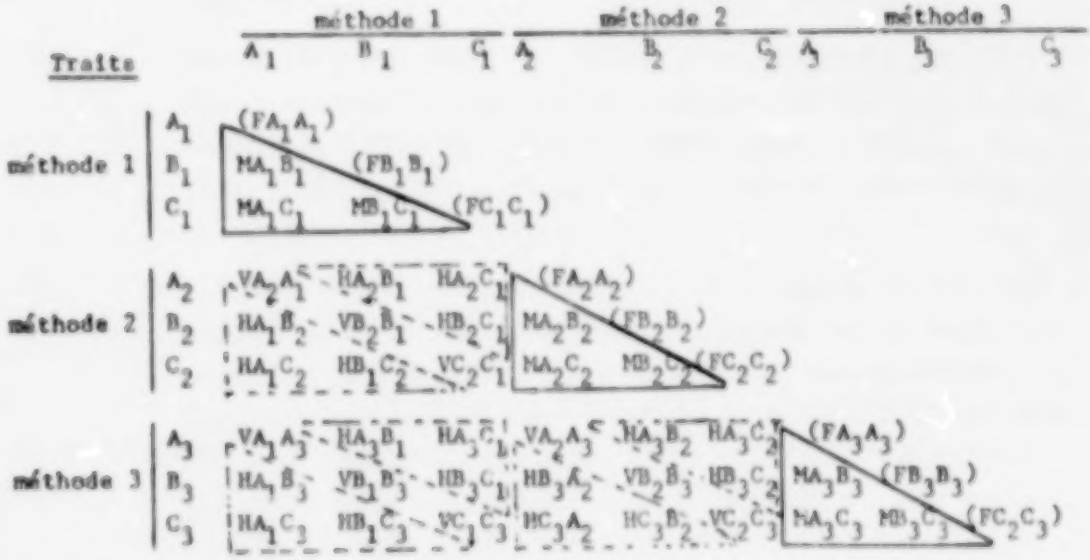
C) Mesures d'insatisfaction

Nous abordons cette mesure avant les mesures A et B parce que c'est le construit le plus utilisé dans les résultats au test PERPE. Ce construit est basé sur une conception de l'insatisfaction exprimée en termes de différence entre la perception d'une réalité et d'une attente. Une absence de différence devient alors indicative de satisfaction. Une telle approche n'est pas nouvelle: Porter (voir Dubois (17)) l'a déjà opérationnalisée à partir des travaux de Rosen et Rosen, Dubois et al.(17) l'ont utilisée au Québec; Smith et all. (47) ont fait de même chez nos

voisins du sud. Levinthal et al. (35) étudient cette approche et encouragent son développement. Chabot (12) justifie cette approche par certaines théories psychologiques. Il semble donc utile de considérer l'insatisfaction de cette façon.

Le processus de validation, tel que défini dans notre première partie, accordait une importance primordiale à la validité convergente et discriminante d'un instrument. Notre instrument sera en effet précieux en autant qu'il pourra discriminer l'insatisfaction étudiante sur différents domaines tels l'encadrement, le climat... (validité discriminante). Cette discrimination est importante puisque, sans elle, on ne pourrait différencier l'insatisfaction envers le climat de celle envers la disponibilité; notre instrument s'avérera d'autre part précis s'il mesure l'insatisfaction conformément à celle mesurée par d'autres méthodologies équivalentes. Il est temps maintenant d'expliciter cette méthodologie initiée par Campbell et Fiske (10).

Cette méthodologie est basée sur des comparaisons de corrélations entre des mesures de plusieurs traits par plusieurs méthodes de mesures, d'où son nom "multitrait-multimethod matrix". La figure suivante illustre cette matrice.



Supposons que les trois traits mesurés sont le climat, la versatilité et le magnétisme et que les trois méthodes sont la cote MAL, la cote MAB et un interview.

Les corrélations inclues dans les triangles solides représentent des corrélations entre des traits différents mesurés par une même méthode; on les appelle "hétérotrait-heteromethod blocks".

Les corrélations diagonales V représentent des corrélations entre un même trait mesuré par différentes méthodes; on les appelle "monotrait-heteromethod blocks".

Les indices F, entre parenthèses, représentent un estimé de la fidélité d'un trait mesuré par une méthode.

L'analyse effectuée sur cette matrice doit conduire à la vérification des hypothèses suivantes:

1. Les méthodes de mesure ne doivent pas affecter le lien entre des mesures d'un même trait. Donc, les corrélations V doivent être au moins statistiquement significative et habituellement plus élevées que les autres corrélations de la matrice; ces corrélations sont des indices de validité convergente.

2. Les corrélations entre les mesures d'un même trait par différentes méthodes doivent être plus élevées que celles entre les mesures de differents traits par différentes méthodes. Donc, $VA_2A_1 > HA_2B_1$, HA_2C_1, HA_1B_2, HA_1C_2. C'est le premier élément de validité discriminante.

3. Les corrélations entre différents traits mesurés par une même méthode ne doivent pas être plus élevées que celles entre le même trait mesuré par différentes méthodes, car on serait tenté d'attribuer cette corrélation à la méthode utilisée plutôt qu'au trait mesuré. Donc, les corrélations $M \leq V$. Cependant l'effet des méthodes est ici confondu avec la nature du lien entre les traits mesurés. C'est le deuxième élément de validité discriminante.

4. Le "pattern" des corrélations dans les différents triangles (solides et pointillés) doit être le même. Ceci indiquerait l'existence de "corrélations vraies" entre les traits quelles que soient les méthodes utilisées. Donc, la grandeur relative des M et H doit être constante d'un triangle à l'autre. C'est le dernier élément de validité discriminante.

Althouser et al. (1) ont montré que ce modèle ne résistait pas à une approche causale. Ils proposent de réduire cette matrice à deux traits et deux méthodes de mesure et de tenir compte de l'antériorité d'une mesure par rapport à l'autre. La figure suivante illustre la matrice proposée.

	trait	méthode 1	méthode 2
mesure initiale	X_1	X_1 Y_1	X_2 Y_2
	Y_1	rY_1X_1	rX_1X_2 rX_1Y_2
			rY_1X_2 rY_1Y_2
mesure subséquente	X_2		
	Y_2		rY_2X_2

Nous pouvons ainsi étudier les corrélations entre deux méthodes de mesure d'un même trait (rX_1X_2, rY_1Y_2), entre deux méthodes de mesure de différents traits (rX_1Y_2, rY_1X_2) et entre les mesures successives de différents concepts (rY_1X_1, rY_2X_2). Cette nouvelle matrice exigerait l'approfondissement du modèle causal présenté par les auteurs; je vois l'application de ce schème au niveau des mesures A et B.

Smith et al. (47), dans leur effort de valider un questionnaire de satisfaction au travail, présente une stratégie originale construite à partir des hypothèses de Campbell et Fiske (10). Sa méthodologie est celle que nous proposons pour valider les mesures de l'ERPE.

Ces auteurs, assumant que la satisfaction sur différents aspects est une perception subjective, valident la mesure qu'ils utilisent avec d'autres mesures de cette perception subjective; ils considèrent valides les méthodes de mesure qui apparaissent consistantes pour mesurer chaque aspect et discriminantes dans leurs mesures des différents aspects. Utilisant l'analyse de "cluster" et l'analyse factorielle, ils évalueront la validité convergente et discriminante de leur méthode de mesure.

ETAPE 1

Pour effectuer cette validation, nous avons besoin de plusieurs traits et plusieurs méthodes. Les traits mesurés pourraient facilement correspondre aux six sections actuelles du test puisque celles-ci furent construites à partir de regroupements empiriques; quoique non indépendants, ces traits mesurent quand même des aspects différents de la situation d'enseignement.

Dans le test actuel, nous avons deux mesures d'insatisfaction, la MAB et la MAL, nous suggérons d'utiliser ces deux mesures afin d'en saisir les relations et de les comparer à deux autres mesures d'insatisfaction. La première consisterait à utiliser une échelle d'insatisfaction, en 8 points, le point 0 correspondant à l'absence d'insatisfaction, le point 7 à une très grande insatisfaction.

aucune insat. — très grande insat.

0 1 2 3 4 5 6 7

Notre deuxième mesure utiliserait une échelle ordinale en trois catégories, une pour la perception de la réalité, l'autre pour la perception d'une situation idéale. Par exemple:

Quantité de matière: description...

Présentement, il y a			Idéalement, il y aurait		
peu	moyennement	beaucoup	peu	moyennement	beaucoup
1. ☐	2. ☐	3. ☐	1. ☐	2. ☐	3. ☐

Deux réponses identiques (1-1, 2-2, 3-3) obtiendraient une cote de 0. Deux réponses différentes obtiendraient une cote de +1 ou -1 selon le sens de l'insatisfaction. Nous obtiendrions ici deux cotes, une d'insatisfaction brute (somme en valeurs absolues) et une d'insatisfaction nette (somme algébrique). Donc, nous avons six traits mesurés et cinq mesures de l'insatisfaction.

ETAPE 2

Chaque groupe d'item mesurant un aspect (section) est divisé en deux sous-groupes selon le pouvoir discriminatif des item pour les deux mesures MAB et MAL du test. Le pouvoir discriminatif réfère simplement au pourcentage d'étudiants exprimant une insatisfaction. Nous avons donc en tout 42 scores, soit 6 sections comprenant deux mesures MAL, deux mesures MAB, un score d'insatisfaction, un score net d'insatisfaction (IPI insatisfaction-présent-idéal) et un score brut d'insatisfaction (IPIA - insatisfaction-présent-idéal-absolu).

ETAPE 3

L'échantillon utilisé est divisé en deux groupes pour des fins de reproductibilité. Ceci permet de confirmer les recherches effectuées sur un groupe par les résultats d'un deuxième groupe.

ETAPE 4

Un "cluster analysis" (voir TRYON (51)) est effectué. Cette méthode permet de situer chaque variable dans un "cluster" seulement si elle est en corrélation plus forte avec une variable déjà incluse dans ce "cluster" qu'avec des variables appartenant à d'autres "cluster". Ces "cluster" peuvent être déterminés "a priori".

ETAPE 5

Une analyse factorielle (principal component method, varimax rotation, with unity in the main diagonal) est ensuite effectuée, le nombre de facteurs étant déterminé par l'analyse des "cluster". Nous suggérons le choix de 8 facteurs, un pour chaque section, un pour les mesures d'insatisfaction brute, l'autre pour les mesures d'insatisfaction nette. Nous obtiendrons un tableau de résultats ayant la forme suivante.

SECTION		Facteurs								h_2
	INSATISFACTION	1	2	3	4	5	6	~	8	
climat	MAB (+ discrim.)									
	MAB (- discrim.)									
	MAL (+ discrim.)									
	MAL (- discrim.)									
	IPI									
	IPIA									
disponibilité										
⋮										
pourcentage de variance commune aux 8 facteurs										
pourcentage de variance totale										

L'analyse de la matrice des corrélations (Campbell et Fiske), du "cluster analysis" et de l'analyse factorielle (Smith et al) nous permettra d'établir clairement la validité convergente et discriminante de nos mesures.

D) Mesures A et B

Il serait intéressant de valider la mesure A avec la norme "présentement" expérimentée dans C, au moyen de la matrice de Campbell et Fiske. Il pourrait aussi être intéressant de construire une grille d'observation sur le modèle proposé par Medley et Mitzel (voir Gage (22)). Ceci nous permettrait de donner un poids empirique à l'hypothèse de Chabot (12) sur la correspondance entre des moyennes de perceptions de réalité et la réalité observée.

Pour la mesure B, une étude de la notion d'attentes et d'idéal pourrait être amorcée par les résultats de l'expérimentation. A ce sujet, une étude empirique des modèles proposés par Lamontagne (31) serait très utile.

E) En conclusion, ce texte ne sert qu'à faire démarrer une démarche de validation sérieuse; il reste à entretenir les rouages du système, à en consolider les pièces et à l'offrir sur le marché. Rappelons que c'est surtout à la lumière des résultats de la validation des mesures d'insatisfaction que des hypothèses pourront être posées clairement au niveau des mesures A et B.

Références et Bibliographies

1. ALTHAUSER, R.P., HEBERLEIN, T.A., SCOTT, R.A. (1971). A causal assessment of validité: the augmented multitrait-multimethod matrix in H.M. Blalock (ed.) Causal models in the social sciences. Aldine Atherton pp. 374-399.

2. AMERICAN PSYCHOLOGICAL ASSOCIATION (1966). Standards for eductional and psychological tests and manuals. Washington, D.C.

3. AMERICAN PSYCHOLOGICAL ASSOCIATION, Committee on Psychological Tests (1954). Technical Recommendations for psychological tests and diagnostic techniques. Washington, D.C.

4. ANASTASI, Anne (1968). Psychological Testing (3e édition). New York, McMillan.

5. BALMA, M.J. (1959). The concept of synthetic validity. Personnel Psychology, 12, 395-396.

6. BARR, A.S. et al. (1961). Wisconsin studies of the measurement and prediction of teacher effectiveness. 4e ed. (1968). Dembar publications, Inc. Madison, Wisconsin 53703.

7. BECHTOLDT, H.P. (1959) Construct validity: a critique. American Psychologist, 14, 619-629.

8. BLAKE et MOUTON

9. CAMPBELL, D.T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. American Psychologist, 15, 546-553.

10. CAMPBELL, D.T. & FISKE, D.W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, 56, 81-105.

11. CATTELL, R.B. (1964). Validity and reliability: a proposed more basic set of concepts. Journal of Educational Psychology, 55, 1-22.

12. CHABOT, Marie (1970). La validité du test PERPE. PERPE, document 0570-02.

13. CHABOT, Marie & GANGE, F. (1970) PERPE, une conception neuve du perfectionnement pédagogique. Prospectives, 6, no. 3, pp. 160-181.

14. COSTIN, F. (1971). Empirical test of the "teacher-centered" versus "student-centered" dichotomy. Journal of Educational Psychology, 62, 410-412.

15. CRONBACH, L.J. & MEEHL, P.E. (1955). Construct validity in psychological tests. Psychological bulletin, 52, 281-302.

16. CURETON, E.E. (1951) Validity in E.F. Lindquist (ed.) Educational Measurement. American Council on Education, Washington, D.C.pp.621-94.

17. DUBOIS, P, & ARCHAMBAULT, J. & BOISSONNEAULT, R. (1969). La société face au crime, annexe 8: satisfaction au travail des policiers municipaux du Québec. Editeur officile du Québec, Gouvernement du Québec.

18. EBEL, R.L. (1956). Obtaining and reporting evidence on content validity. Eductional and Psychological Measurement, 16, 269-82.

19. EBEL, R.L. (1961) Must all tests be valid? American Psychologist, 16, 640-647.

20. FOUCHER, R. (1972) La satisfaction des étudiants: une composante importante de l'apprentissage? Prospectives, 8, no. 2, pp. 67-73.

21. FOX,A.M., BROOKSHIRE, W.K. (1971) Défining effective college teaching. Journal of experimental education, 40, 37-40.

22. GAGE, N.L. (1961) Paradigms for research on teaching in Gage, N.L. (ed.) Handbook of research on teaching. Rand McNally cie, Chicago, 94-141.

23. GAGNE, F. (1969) Qu'est-ce que le projet PERPE? PERPE, doc. 1169-02.

24. GAGNE, F. (1971) 33,000 répondants évaluent la pédagogie au niveau collégial. PERPE , document 1071-04.

25. GAGNE, F. (1972) Perceptions étudiantes de l'enseignement collégial in L'étudiant québécois: Défi et dilemnes. Ministère de l'éducation Gouvernement du Québec. PP 113-180.

26. GAGNON, Denise (1971). Historique et caractéristiques psychométriques du test PERPE. PERPE, document 0671-05.

27. GOODENOUGH, F.L. (1949). Mental testing: its history, principles and applications. New York: Rinehart.

28. GULLIKSEN, H. (1950). Intrinsic validity. American Psychologist, 5, 511-517.

29. HAMACHEK, D.E. (1968) What research tells us about the characteristics of "good" and "bad" teachers , in D.E.Hamachek (ed.) Human dynamics in psychology and education. Boston, Allyn and Bacon, pp. 187-203.

30. LAMONTAGNE, J. (1971) L'analyse des déterminants de la situation pédagogique à partir des données de PERPE. PERPE, document 0671-01.

31. LAMONTAGNE, J. (1972) La centration sur l'étud[illegible] dans la situation éducative. Document inédit. PERPE.

32. LAWSHE, C.H. (1952) Employee selection. Personnal Psychology, 5, 31-34.

33. LAWSHE, C.H. & BALMA, M.J. (1966). Principles of personnel testing. (end ed.) McGraw-Hill, New York.

34. LENNON, R.T. (1956) Assumptions underlying the use of content validity. Educational and Psychological Measurement, 16, 294-304.

35. LEVINTHAL, C.F., LANSKY, L.M. & ANDREWS, O.E. (1971) Student evaluations of teacher behaviors as estimations of real-ideal discrepancies: a critique of teacher rating methods. Journal of Educational Psychology, 62, 104-109.

36. LORD, F.A. (1952) Theory of test scores. Psychometric Monograph, no. 7.

37. LORD, F.M. & NOVICK, M.R. (1968) Statistical theories of mental test scores. Addison-Wesley Publishing Company.

38. MARX, M.H. (?) The General nature of Theory Construction. Edité dans un handbook, référence égarée.

39. MCQUITTY, L.L. (1957) Elementary linkage analysis for isolating orthogonal and oblique types and typal relevance. Educational and Psychological Measurement, 17, 207-229.

40. MEDLEY, D.M. MITZEL, H.E. (1958) Technique for measuring classroom behavior. Journal of Educational Psychology, 49, 86-92.

41. Ministère de l'Education (1971) Le système de probation des enseignants. Ministère de l'Education, Québec.

42. NUNNALLY, J.C. (1967) Psychometric Theory. McGraw-Hill, New York.

43. PERPE (1971) Plan triennal de développement du Service PERPE. En collaboration, PERPE, document 0171-01.

44. PERPE (1972). CRASE: orientations générales et programme d'activités. En collaboration, PERPE, document 0172-01.

45. RYANS, D.C. (1963) Assessment of teacher behavior and instruction. Review of educational research, 33, 415-441.

46. SIMON, H.A. & NEWELL, A. (1956) Models: their uses and limitations in L.D. White (ed.) The state of social science. Chicago, University of Chicago Press, pp. 66-83.

47. SMITH, P.C., KENDALL, L.M. & HULIN, C.L. (1969). The measurement of satisfaction in work and retirement. Rand McNally & cie, Chicago

48. SOLOMON, D. et al. (1964). Dimensions of teacher behavior. Journal of experimental education, 33, 23-40.

49. STEVENS, S.S. (1939). Operationism and logical positivism. Psychological bulletin, 36, 221-263.

51. TRYON (1971) cluster analysis McGraw-Hill